KANTIPUR ENGINEERING COLLEGE

(Affiliated to Tribhuvan University)

Dhapakhel, Lalitpur

[Subject Code: CT654]

A MINOR PROJECT FINAL REPORT ON

HOUSE PRICE PREDICTION SYSTEM USING MULTIPLE LINEAR REGRESSION

Submitted by:

Aman Devkota [KAN076BCT010]

Ankur Karmacharya [KAN076BCT013]

Prashad Adhikary [KAN076BCT056]

A MINOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

March, 2023

ABSTRACT

The growth of Machine learning has been rapid in this past decade. Many applications evolve in Machine learning day by day. One such application is the house price prediction. Humans are very thoughtful when they want to make investments in a house. The prices for houses have been changing every year that has necessitated the modelling of a house price prediction system. This system will make use of the features of the house such as number of bathrooms available, total area, location of the house etc. to generate an estimated price for the house with the help of multiple linear regression.

Keywords *– multiple linear regression, house price prediction system*

ACKNOWLEDGMENT

We would like to express sincere gratitude to Department head Er. Rabindra Khati, Project Co-ordinator Er. Bishal Thapa and all the faculty members of Kantipur Engineering College for the continuous support during this project for their patience, motivation,enthusiasm, and immense knowledge. Their guidance helped us in all time of research, development and implementation of this project.

Aman Devkota [KAN076BCT010]
Ankur Karmacharya [KAN076BCT013]
Prashad Adhikary [KAN076BCT056]

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Usually when people want to buy a house, they look for a house which has a reasonable cost, and which has all the desired features they want in the house. The house price prediction will help them to decide whether the house they desire to buy is worth of the price or not. Similar is the case with people who want to sell the house. By making use of the house price prediction system, the seller would be able to decide what all features he/she could add in the house so that the house can be sold for a higher price. [1]

Modelling uses machine learning algorithms, where machine learns from the data and uses them to predict a new data. The most frequently used model for predictive analysis is regression. As we know, the proposed model for accurately predicting future outcomes has applications in economics, business, banking sector, health care industry, e-commerce, entertainment, sports etc. One such method used to forecast house prices are based on multiple factors. In metropolitan cities, the prospective home buyer considers several factors such as location, size of the land, road conditions, parking availability and most importantly the house price. Multiple linear regression is one of the statistical techniques for assessing the relationship between the (dependent) target variable and several independent variables. Regression techniques are widely used to build a model based on several factors to predict price. [2]

Multiple linear regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple linear regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

The dataset that we are using contains the data about houses in Bangalore from the year 2019. It contains the following attributes:
Area: Area of the property in square feet
BHK: Number of bedrooms along with 1 hall and 1 kitchen

Bathrooms: Number of bathrooms

Location: Location in which property lies

Price: This is the Price of property in INR

Area Type: It's an Apartment or Builder Floor

1.2 Problem Definition

Prices of real estate properties are sophisticatedly linked with our economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available. Therefore, the goal of this project is to use machine learning to predict the selling prices of houses based on many factors.

1.3 Objectives

The primary objectives of this projects are as follows:

i. To predict house prices on the basis of various parameters.

1.4 Project Features

The project will be able to accomplish following:

- Low cost
- Accurate
- User friendly

1.5 Application Scope

The application scope of our project is in real estate business. Our project will allow the buyer to get an idea of what amount of money he/she has to spend in order to buy the desired house. It will also allow the seller to get information regarding what the estimated worth of house is and how he/she can maximize the profit gained by selling the house.

1.6 System Requirement

1.6.1 Software Requirements

Windows/Linux/Mac

HTML/CSS/JS

Jupyter Notebook

Python IDE

1.6.2 Hardware Requirements

PC with at least 4-8 GB RAM

Higher graphics of at least 2 GB

1.7 Project Feasibility

1.7.1 Technical Feasibility

The technical feasibility assessment is focused on gaining in understanding of the present technical resources required by the system and their applicability to the expected needs of the proposed system. Regarding the proposed system, the technical requirement includes a PC.

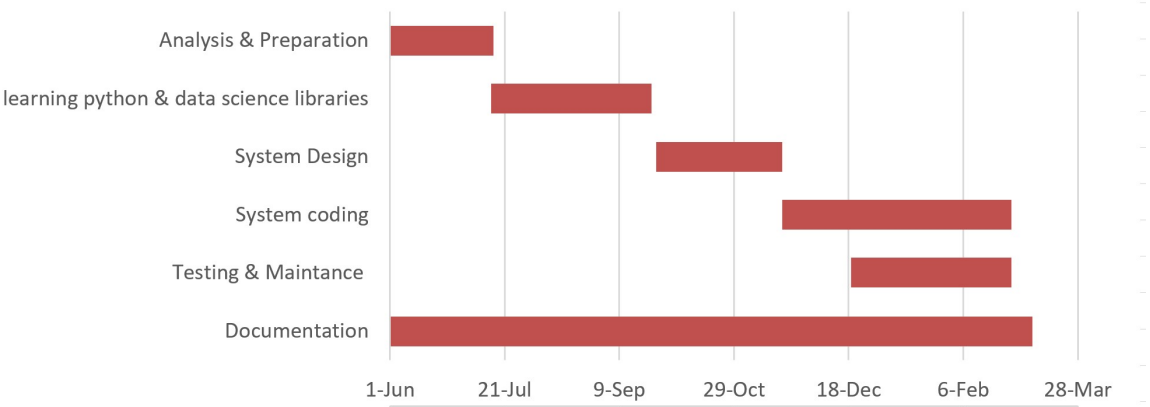
1.7.2 Operational Feasibility

The user will not need any formal knowledge about programming so our project is operationally feasible.

1.7.3 Economic Feasibility

The purpose of the economic feasibility assessment is to determine the positive economic benefits to the user that the proposed system will provide. Most of the software used for the development is free. Thus, the project is economically feasible.

1.7.4 Schedule Feasibility



	Documentation	Testing & Maintance	System coding	System Design	learning python & data science libraries	Analysis & Preparation
start date	1-Jun	19-Dec	19-Nov	25-Sep	15-Jul	1-Jun
estimated days of completion	280	70	100	55	70	45

Figure 1.1: Gantt Chart

CHAPTER 2
LITERATURE REVIEW

2.1 Related Projects

2.1.1 Zillow

Zillow is a popular online real estate marketplace that provides an automated valuation model called the Zestimate. This model uses a combination of data from public records, user-submitted data, and machine learning algorithms to predict the value of a home.

2.1.2 Redfin's Home Value Estimator

Redfin is another online real estate marketplace that provides a home value estimator tool that uses machine learning algorithms to predict the value of a home based on its location, features, and recent sales in the area.

2.1.3 House Canary

HouseCanary is a real estate analytics company that provides a range of services, including home value estimates and forecasts, neighborhood insights, and market analytics. Their home value estimates are based on a proprietary machine learning model that analyzes millions of data points.

2.1.4 Propmix.io

PropMix.io is a real estate data and analytics platform that offers a range of tools for real estate professionals, including a home value estimator that uses machine learning algorithms to predict the value of a home based on its features and location.

2.2 Related Works

Anirudh Kaushal and Achyut Shankar researched in detail about house price prediction using multiple linear regression method. In the paper "House Price Prediction Using Multiple Linear Regression" published on April 25, 2021 there is explanation about filtering of data set, data processing, training and evaluating multiple linear regression model. [1] Manasa, J., Gupta, R., & Narahari, N. S. studied and compared the algorithms for estimation of price of houses in city of Bengaluru in the paper "Machine learning based predicting house prices using regression techniques". [2] M. Thamarai, S P. Malarvizhi have compared the decision tree and regression algorithms in the paper

"House Price Prediction Modeling Using Machine Learning". The basis of comparison was accuracy, MAE, MSE and RMSE. [3] Rana, V. S., Mondal, J., Sharma, A., & Kashyap, I. have studied in detail about the machine learning algorithms and compared the results obtained by each to learn about the algorithm most suitable to use in the house price prediction system in their paper "House Price Prediction Using Optimal Regression Techniques".[4] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. studied on different regression algorithms. In the paper "House price prediction using regression techniques: a comparative study" published on March, 2019, there is explanation about different types of regression methods and their accuracy to predict the values. [5]

CHAPTER 3
METHODOLOGY

3.1 Working Mechanism

The development of house price prediction system involves major steps which is depicted in the diagram given below:

3.1.1 Load the data set

The raw data of different houses with different parameters as independent values and the price of the house as dependent value are collected and are loaded for data analysis.

3.1.2 Add dummy variables

The parameters in the data set like addresses are string values but the multiple linear regression can only accept numerical values so they are need to be converted into numerical values. In order to convert it we need to give them the value 0 and 1 on the basis that they are available but there may arise dummy trap problem (when two independent parameters affect each other which can cause conflict in multiple linear regression). So to solve this problem the dummy variables should be one less than the n number of string valued parameters.

3.1.3 Filter the dataset

All the parameters in the raw data set are not needed for multiple linear regression as some of the parameters have little or no significance in changing the price of the house. So to filter out the useless parameters, we find the single correlation between price of the house and the parameters. Correlation coefficients are used to measure the strength of the linear relationship between two variables. A correlation coefficient greater than zero indicates a positive relationship while a value less than zero signifies a negative

relationship.If the correlation is very low or near to zero, they can be neglected.
Mathematical calculation for correlation

$$(r_{xy}) = \frac{(n\sum xy - \sum x \sum y)}{(\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{3.1}$$

We used Regular expression to filter out the location attributes.

3.1.4 Split the data set into train set and test set

The data set is divided in train set and test set so that the train set can be used for training the model.

3.1.5 Statistical Calculations

For the multiple linear regression, we need to find the values for $\sum x, \sum xy, ... \sum x_n$ etc. which are calculated in this stage.

3.1.6 Representing the normal equation of multiple linear regression in matrix form

The equation of plane is

$$x = w_0 + w_1x_1 + w_2x_2 + ...w_Nx_N \tag{3.2}$$

Here,$x_1, x_2, ...x_N$ are independent parameter variables and x is dependent parameter variable. The intercept is w_0 and coefficients are $w_1, w_2, ...w_N$.
Now to find the values of $w_0, w_1, w_2, ...w_N$, we need the normal equations which are as follows:

$$\sum x = nw_0 + w_1\sum x_1 + w_2\sum x_2 + ...w_N\sum x_N \tag{3.3}$$

$$\sum xx_1 = w_0\sum x_1 + w_1\sum {x_1}^2 + w_2\sum x_1x_2 + ...w_N\sum x_1x_N \tag{3.4}$$

$$\sum xx_2 = w_0\sum x_2 + w_1\sum x_1x_2 + w_2\sum {x_2}^2 + ...w_N\sum x_2x_N \tag{3.5}$$

:

$$\sum xx_N = w_0 \sum x_N + w_1 \sum x_1 x_N + w_2 \sum x_2 x_N + ...w_N \sum {x_N}^2 \qquad (3.6)$$

Python does not recognize equations so we need to represent these equations in matrix form for further calculations.

3.1.7 Calculation of intercept and coefficients using Gauss Elimination

To find the values of $w_0, w_1, w_2, ...w_N$, we need to solve the matrices. For this we use the Gauss Elimination method.

Algorithm:

1. Start
2. Declare the variables and read the order of the matrix N
3. Take the coefficients of the linear equations as:

   Do for k= 1 to n

   Do for j = 1 to n+1

   Read a[k][j]

   End for j

   End for k
4. Do for k= 1 to n-1

   Do for i= k+1 to n

   Do for j= k+1 to n+1

   a[i][j] = a[i][j] -a[i][k]/a[k][k]*a[k][j]

   End for j

   End for i

   End for k
5. Compute x[n] = a[n][n+1]/a[n][n]
6. Do for k= n-1 to 1

   sum = 0

   Do for j = k+1 to n

   sum += a[k][j]*x[j]

   End for j

x[k] = 1/a[k][k] * (a[k][n+1] - sum)

End for k

7. Display the result x[k]
8. Stop

3.1.8 Calculating predicted price

Now the calculated values of intercept and coefficients are inserted in the equation (3.2) to calculate the predicted price.

3.2 System Diagram

3.2.1 Use case diagram

The use case diagram presents a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. Below figure shows the use case diagram of the system.

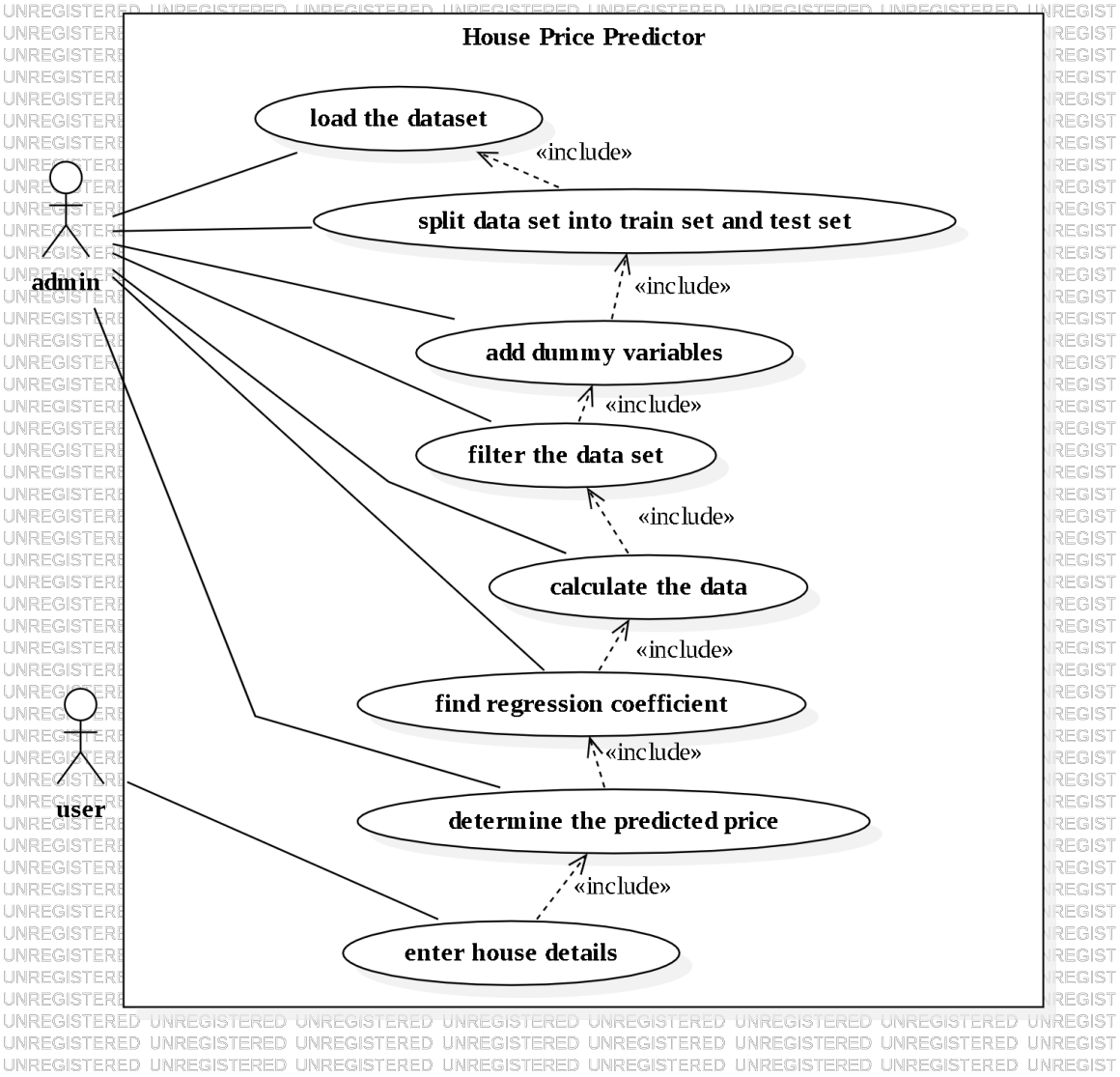


Figure 3.1: Use case Diagram

3.2.2 DFD level diagram

A data flow diagram (DFD) maps out the flow of information for any process or system. In Sanskrit OCR system user feeds Data in form of pictures and images to the system to get digitized document as output.

3.2.3 Software Development Model

Incremental model is a method of software engineering that combines the elements of waterfall model in iterative manner. It involves both development and maintenance.

In this model requirements are broken down into multiple modules. Incremental development is done in steps from analysis design, implementation, testing/verification, maintenance. Each iteration passes through the requirements, design, coding and testing phases. The first increment is often a core product where the necessary requirements are addressed, and the extra features are added in the next increments. The core product is delivered to the client. Once the core product is analyzed by the client, there is plan development for the next increment.
Advantages of Incremental Model:

- Generates working software quickly and early during the software life cycle.
- This model is more flexible – less costly to change scope and requirements.
- It is easier to test and debug during a smaller iteration.
- In this model customer can respond to each built.
- Lowers initial delivery cost.
- Easier to manage risk because risky pieces are identified and handled during it's iteration.

CHAPTER 4
EPILOGUE

4.1 Expected Output

This website will have a good user interface and user experience. In this website, user will have option to either click a picture or upload an image of manuscript/document from camera roll to be digitized. Then this image will go through processes like preprocessing, segmentation, feature extraction, etc. to give the result as shown below. The result can be viewed in the website itself after processing.

4.2 Work Progress and Remaining

4.2.1 Work Completed

We have read research papers regarding the OCR and the algorithms related to it.

• We collected the datasets required for the project

• We trained datasets for the OCR purpose and also tested the data for prediction

• We also completed the segmentation algorithm

So far, the obtained result is as shown in the figure below

4.2.2 Work Remaining

The following are the remaining works:

1. Integration and post processing

REFERENCES

[1] A. Kaushal and A. Shankar, "House price prediction using multiple linear regression," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2021.

[2] J. Manasa, R. Gupta, and N. Narahari, "Machine learning based predicting house prices using regression techniques," in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, pp. 624–630, IEEE, 2020.

[3] M. Thamarai and S. Malarvizhi, "House price prediction modeling using machine learning.," *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 2, 2020.

[4] V. S. Rana, J. Mondal, A. Sharma, and I. Kashyap, "House price prediction using optimal regression techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 203–208, 2020.

[5] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study," in *2019 International conference on smart structures and systems (ICSSS)*, pp. 1–5, IEEE, 2019.

[6] I. Yang *et al.*, "A loss-function for causal machine-learning," *arXiv preprint arXiv:2001.00629*, 2020.